IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tomás Berni
17/12/24

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
- - Data Collection through API
- - Data Collection with Web Scraping
- - Data Wrangling - Exploratory Data Analysis with SQL
- - Exploratory Data Analysis with Data Visualization
- - Interactive Visual Analytics with Folium
- - Machine Learning Prediction
- Summary of all results
- - Exploratory Data Analysis result
- - Interactive analytics in screenshots
- - Predictive Analytics result

Introduction

- ## Project background and context

- SpaceX promotes Falcon 9 rocket launches on its website with a cost of $62 million, which is significantly lower compared to other providers that charge upwards of $165 million. Much of this cost saving is attributed to SpaceX's ability to reuse the first stage. Therefore, if we can accurately predict whether the first stage will successfully land, we can estimate the cost of a launch. This project aims to build a machine learning pipeline that can predict whether the first stage will land successfully. The insights gained from this prediction could be valuable for an alternate company looking to bid against SpaceX for a rocket launch.

- ## Problems you want to find answers

- What factors influence whether a rocket will land successfully?
- The interplay of different elements that impact the likelihood of a successful landing.
- What operating conditions need to be in place to support a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling: One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: How to build, tune, evaluate classification models.

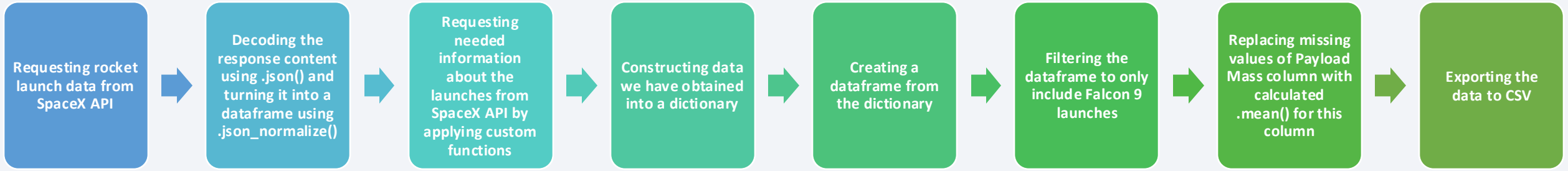
Data Collection

- The data collection process combined API requests from the SpaceX REST API with web scraping of a table from SpaceX's Wikipedia page. Both methods were necessary to gather comprehensive information about the launches, ensuring a more detailed and accurate analysis.
- Data Columns are obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- LINK TO NETBOOK



Data Collection - Scraping

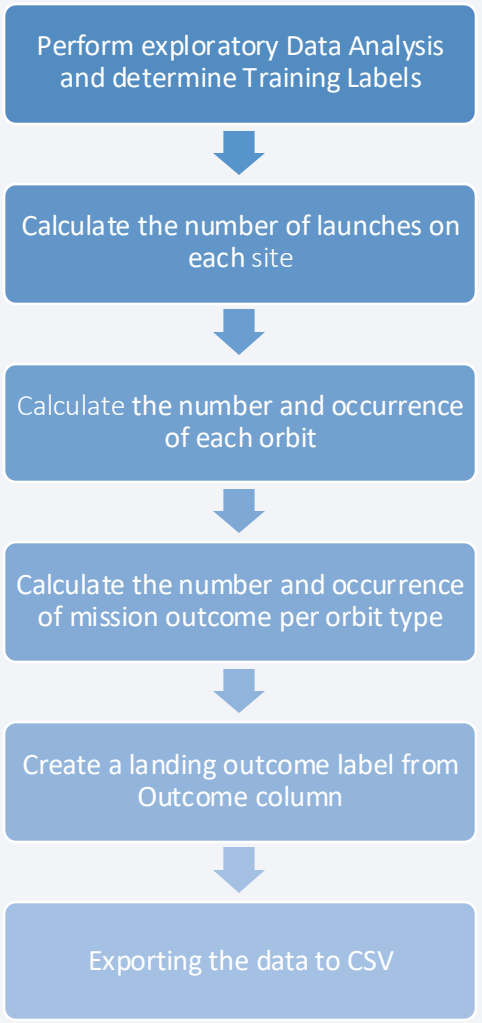
- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup

- LINK TO NOTEBOOK

Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv

- LINK TO NOTEBOOK



EDA with Data Visualization

We created various charts, including:

- Scatter plots to explore the relationships between variables such as Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, and Orbit Type vs. Success Rate. These visualizations can help identify potential relationships that could be valuable for building machine learning models.
- Bar charts for comparing discrete categories. The goal here is to show the relationship between specific categories and a measured value.
- Line charts to illustrate trends over time (time series), such as the yearly success rate trend. These charts provide a clear picture of how data changes over time.

- LINK TO NOTEBOOK

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

- LINK TO NOTEBOOK

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

LINK TO NOTEBOOK

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

- LINK TO NOTEBOOK

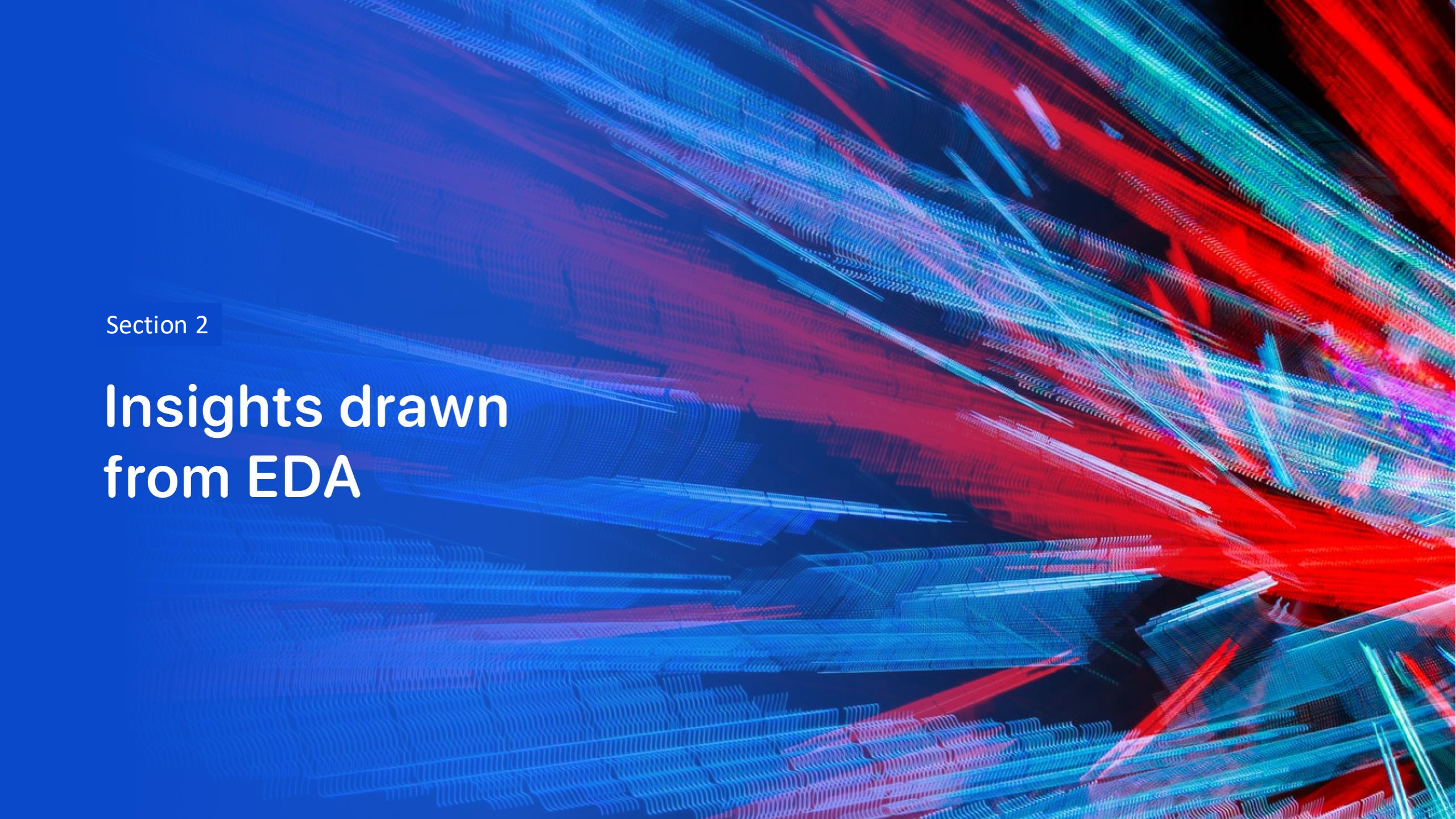
Predictive Analysis (Classification)

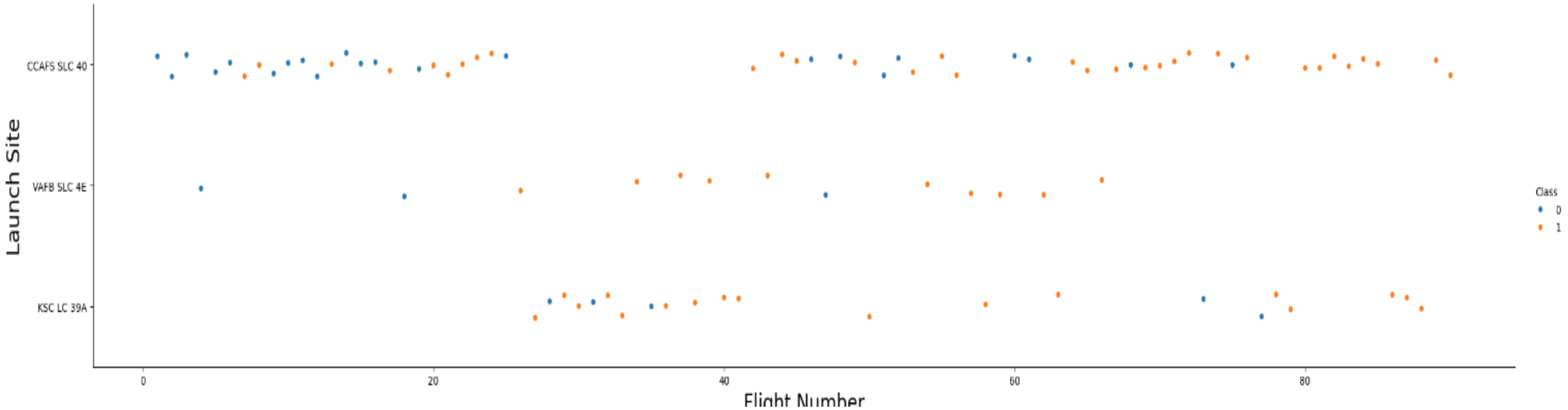
- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.

- LINK TO NOTEBOOK

RESULTS

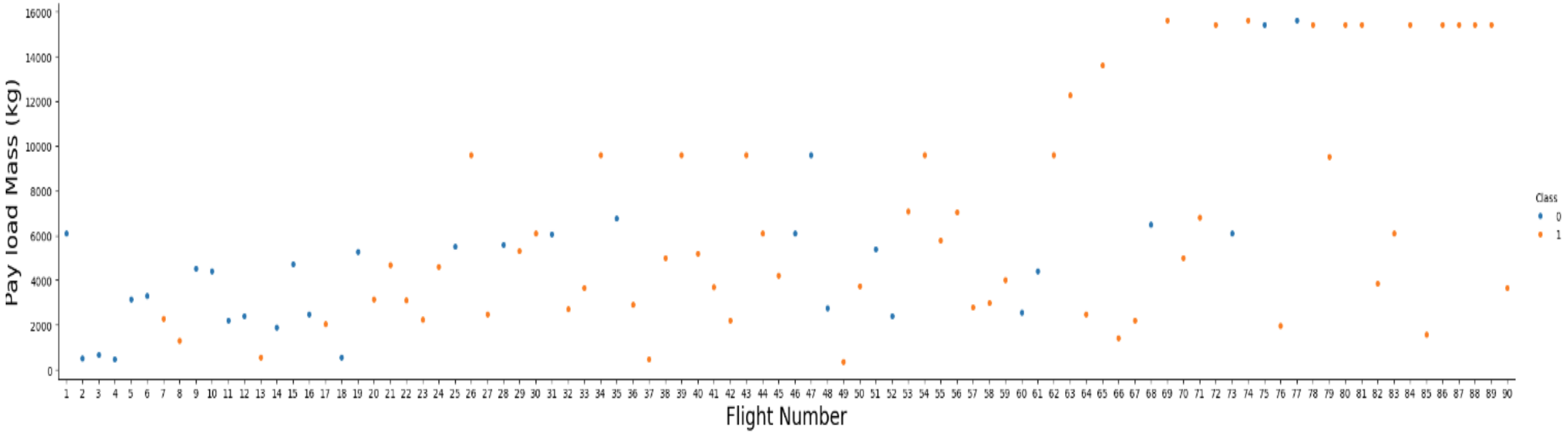
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- **Explanation**:
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.



Payload vs. Launch Site

Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

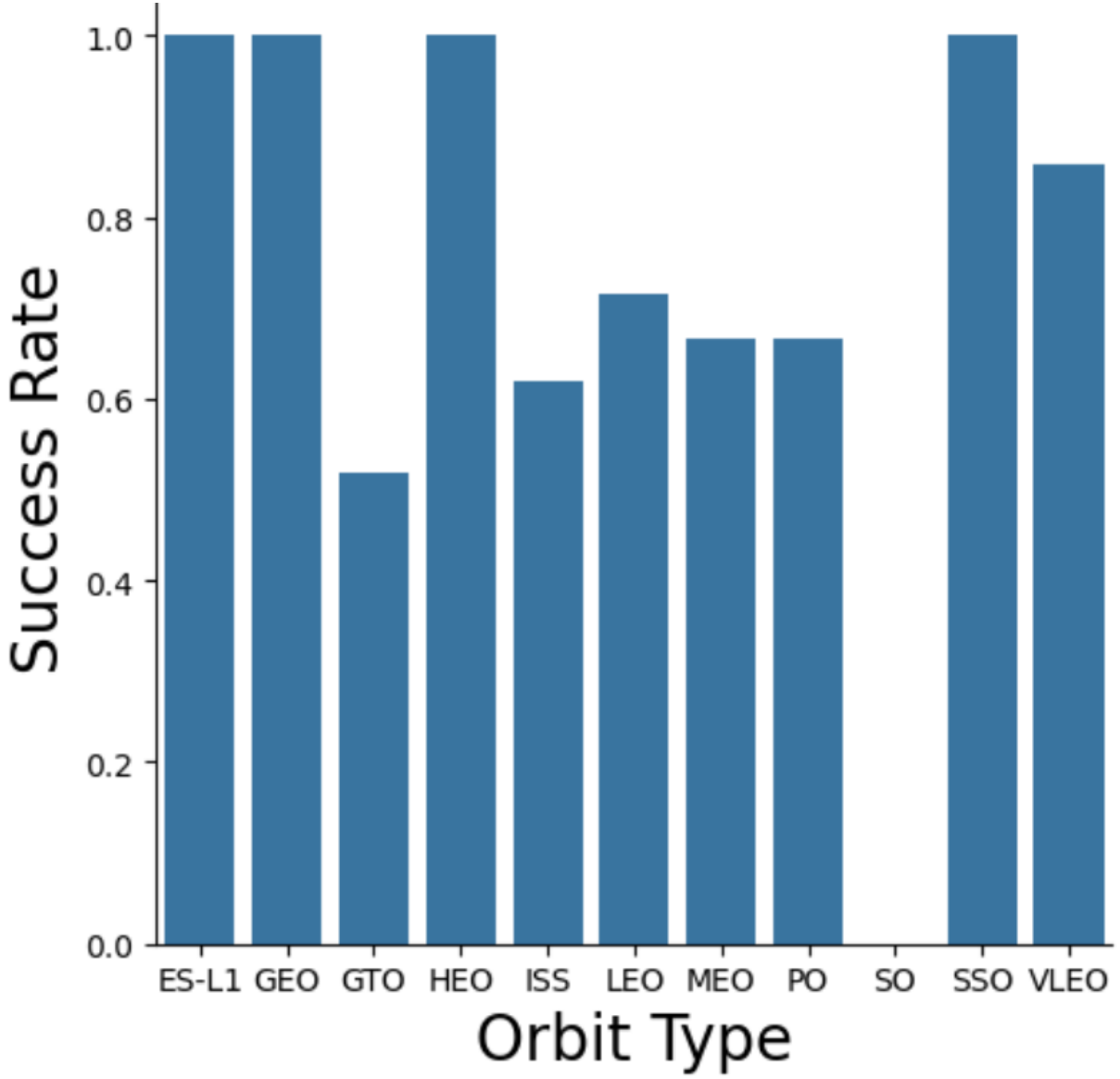
Success Rate vs. Orbit Type

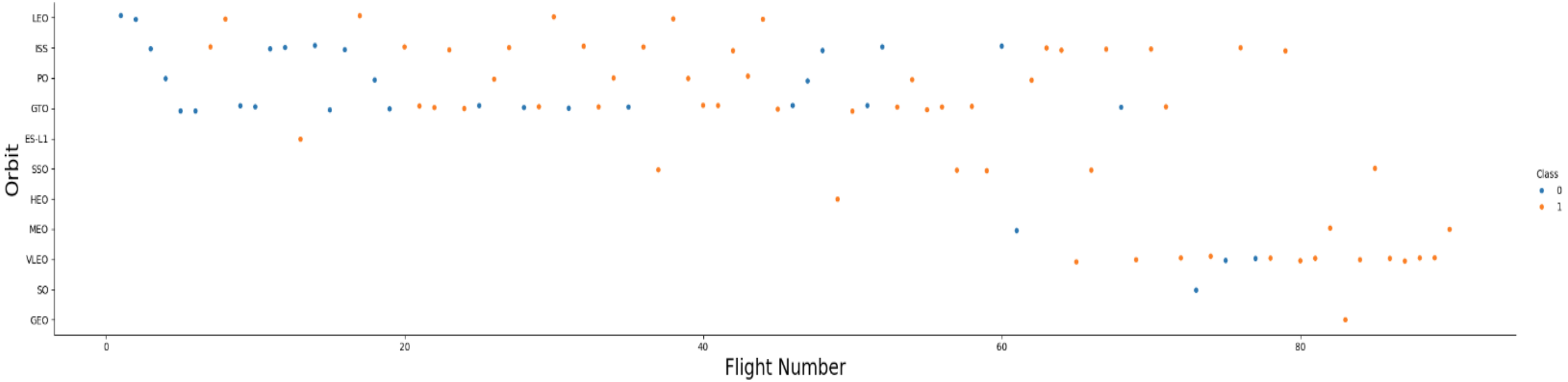
- **Explanation:**

- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO

- Orbits with 0% success rate: - SO

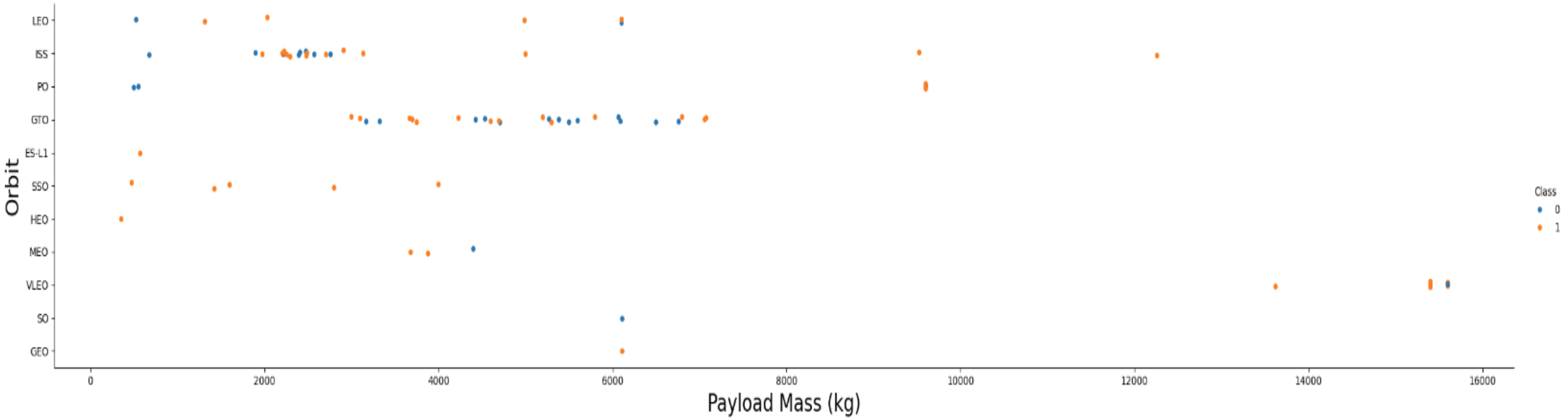
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO





Flight Number vs. Orbit Type

- **Explanation:**
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

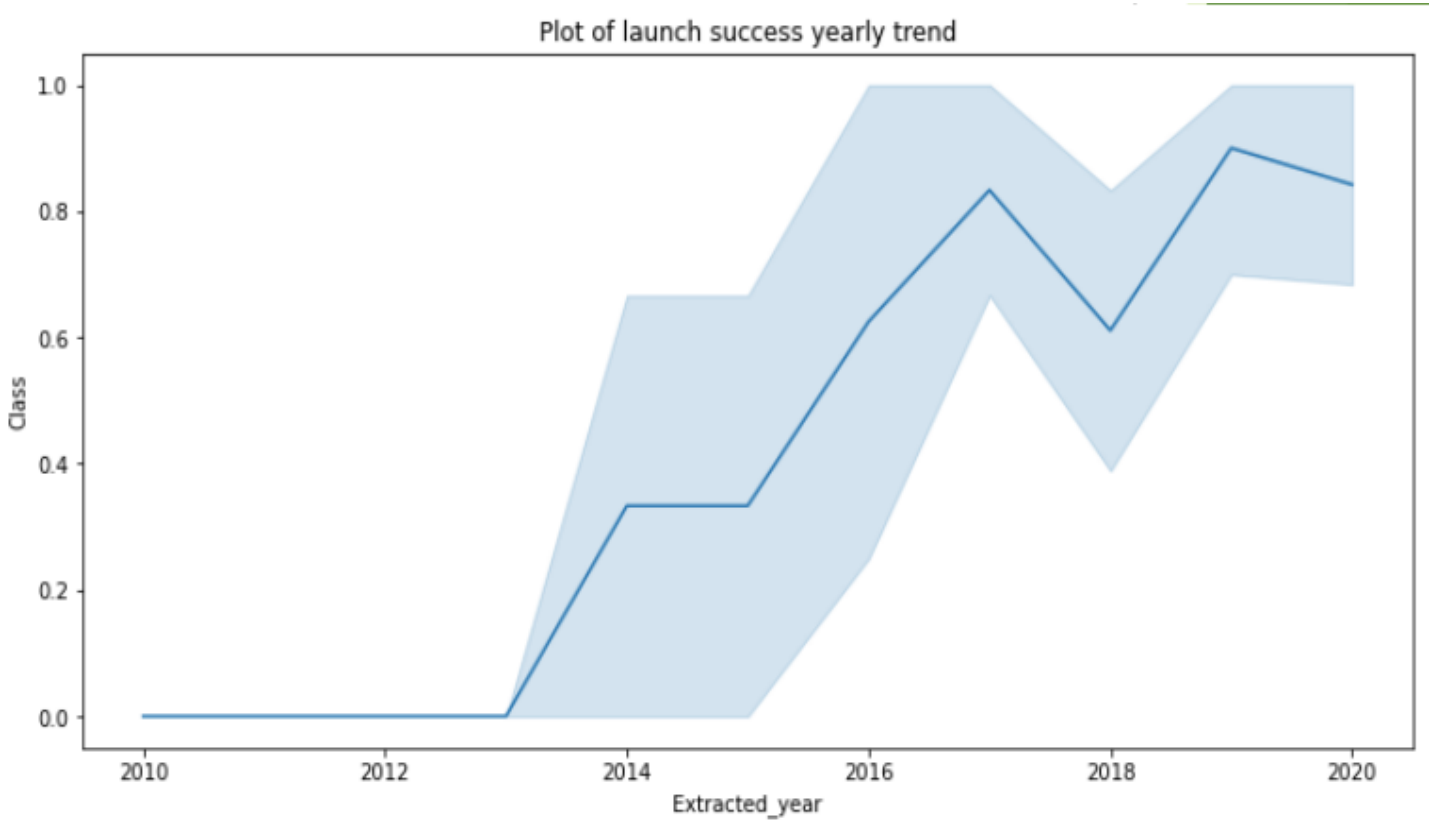


Payload vs. Orbit Type

- **Explanation:**

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- **Explanation:**
- The success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXTABLE;
```

```
* sqlite:///my_data1.db
one.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

Launch Site Names Begin with 'CCA'

- We used the query above to display 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

- **Explanation:**
- Displaying the total payload mass carried by boosters launched by NASA (CRS).

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_versior
```

```
 * sqlite:///my_data1.db
Done.
```

average_payload_mass
2534.6666666666665

Average Payload Mass by F9 v1.1

- **Explanation:**
- Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

- **Explanation:**
- Listing the date when the first successful landing outcome in ground pad was achieved.

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Succe
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Explanation:**

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and p
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- **Explanation:**
- Listing the total number of successful and failure mission outcomes.

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Boosters Carried Maximum Payload

- **Explanation:**

- Listing the names of the booster versions which have carried the maximum payload mass

```
%%sql SELECT strftime('%m', date) AS month, date, booster_version, launch_site, landing_outcome
FROM SPACEXTABLE
WHERE landing_outcome = 'Failure (drone ship)' AND strftime('%Y', date) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

2015 Launch Records

- **Explanation:**
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- **Explanation:**

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

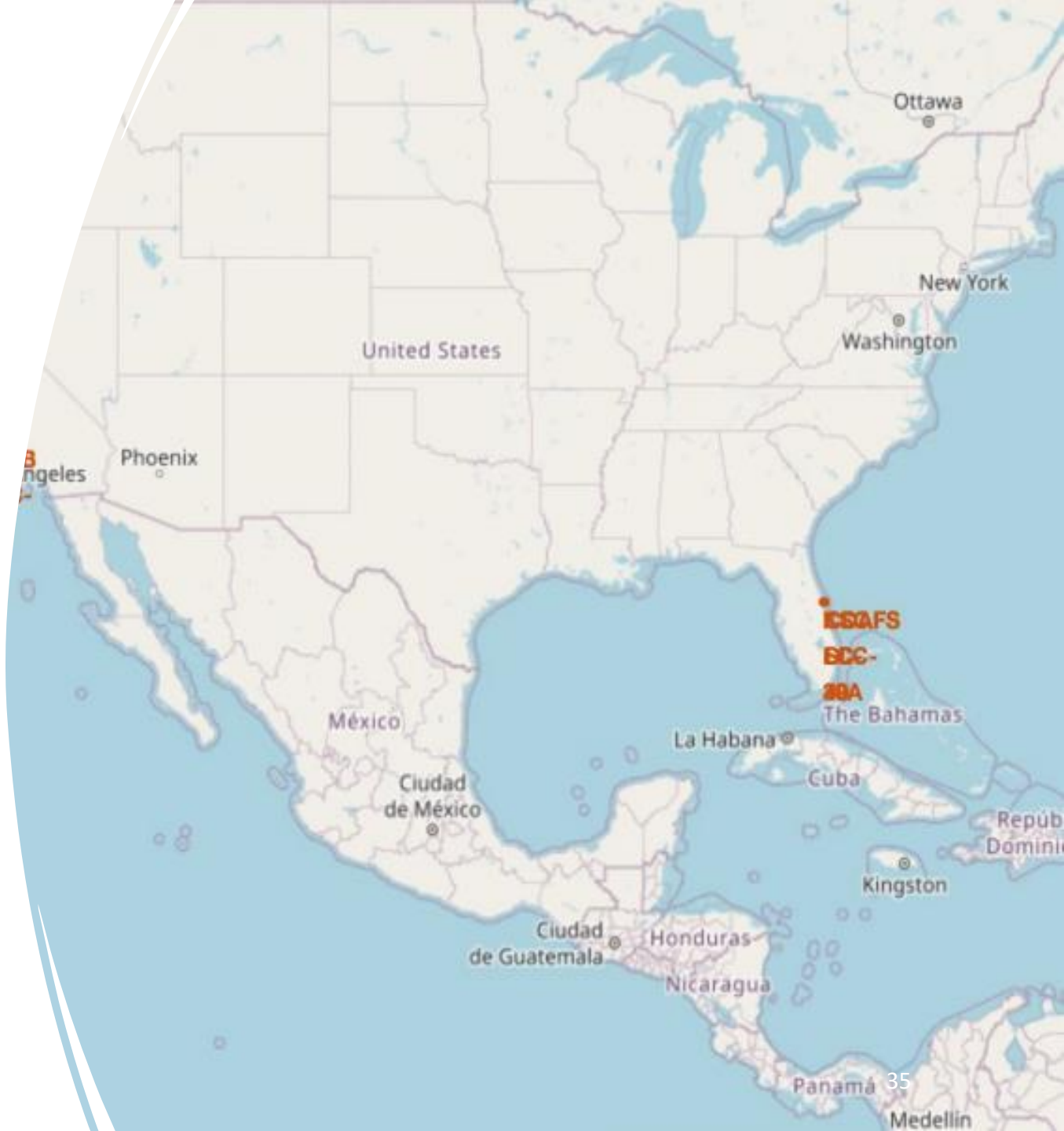
Launch Sites Proximities Analysis

Launch site locations in global MAP

- **Explanation**:

• Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

• All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Markers showing launch sites with color labels

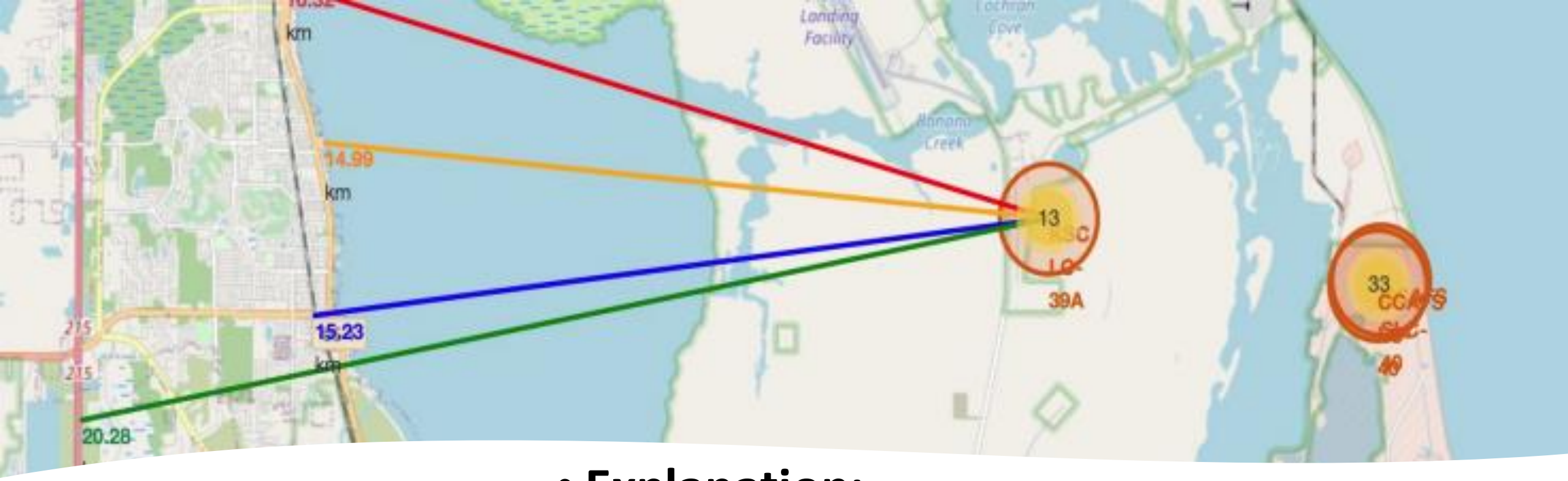
• Explanation:

• From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch

- Red Marker = Failed Launch

• Launch Site KSC LC-39A has a very high Success Rate



Launch Site distance to landmarks

• Explanation:

• From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

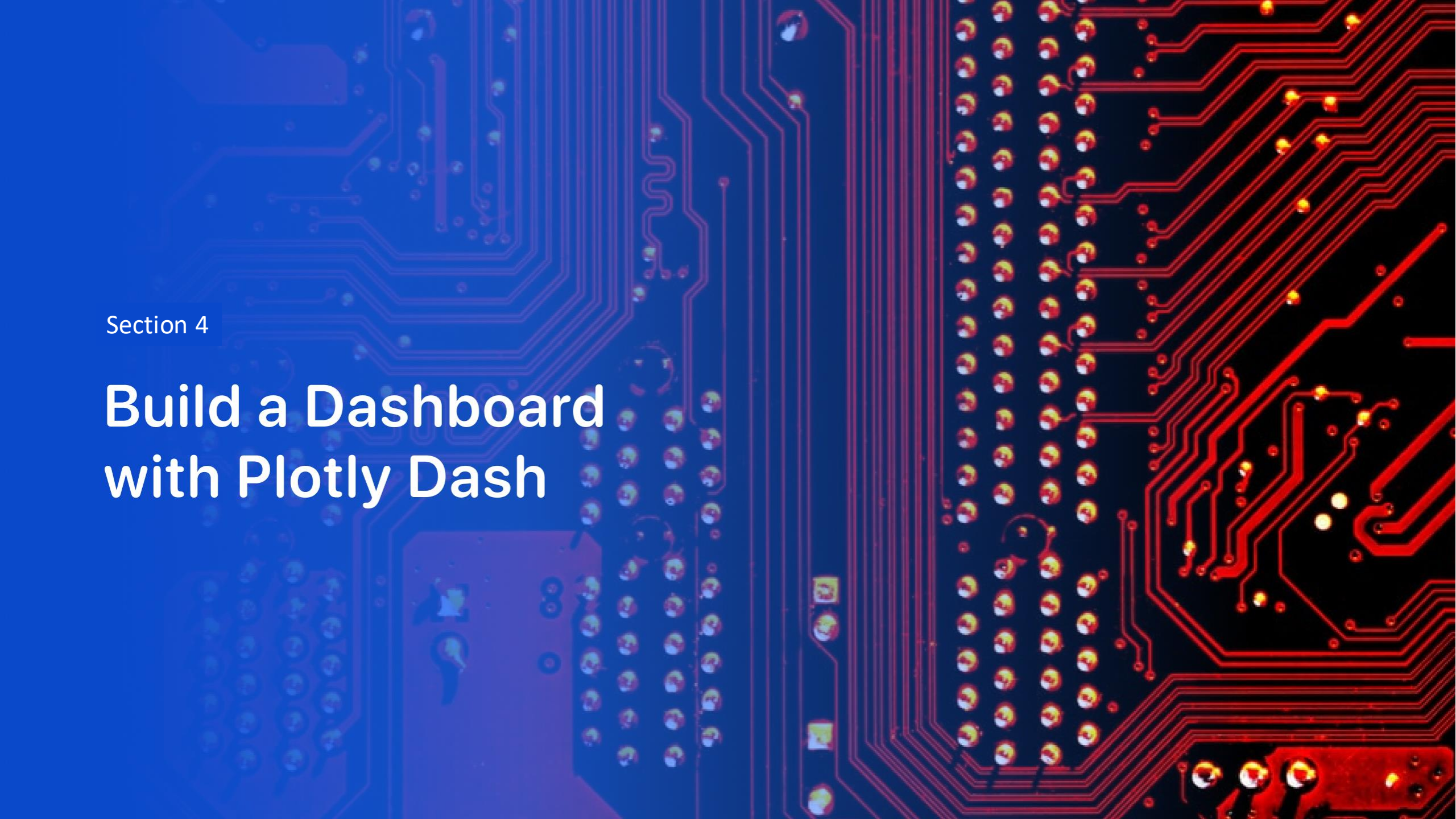
- relative close to railway (15.23 km)

- relative close to highway (20.28 km)

- relative close to coastline (14.99 km)

• Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

• Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.


Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

- **Explanation:**

• The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Pie chart showing the Launch site with the highest launch success ratio

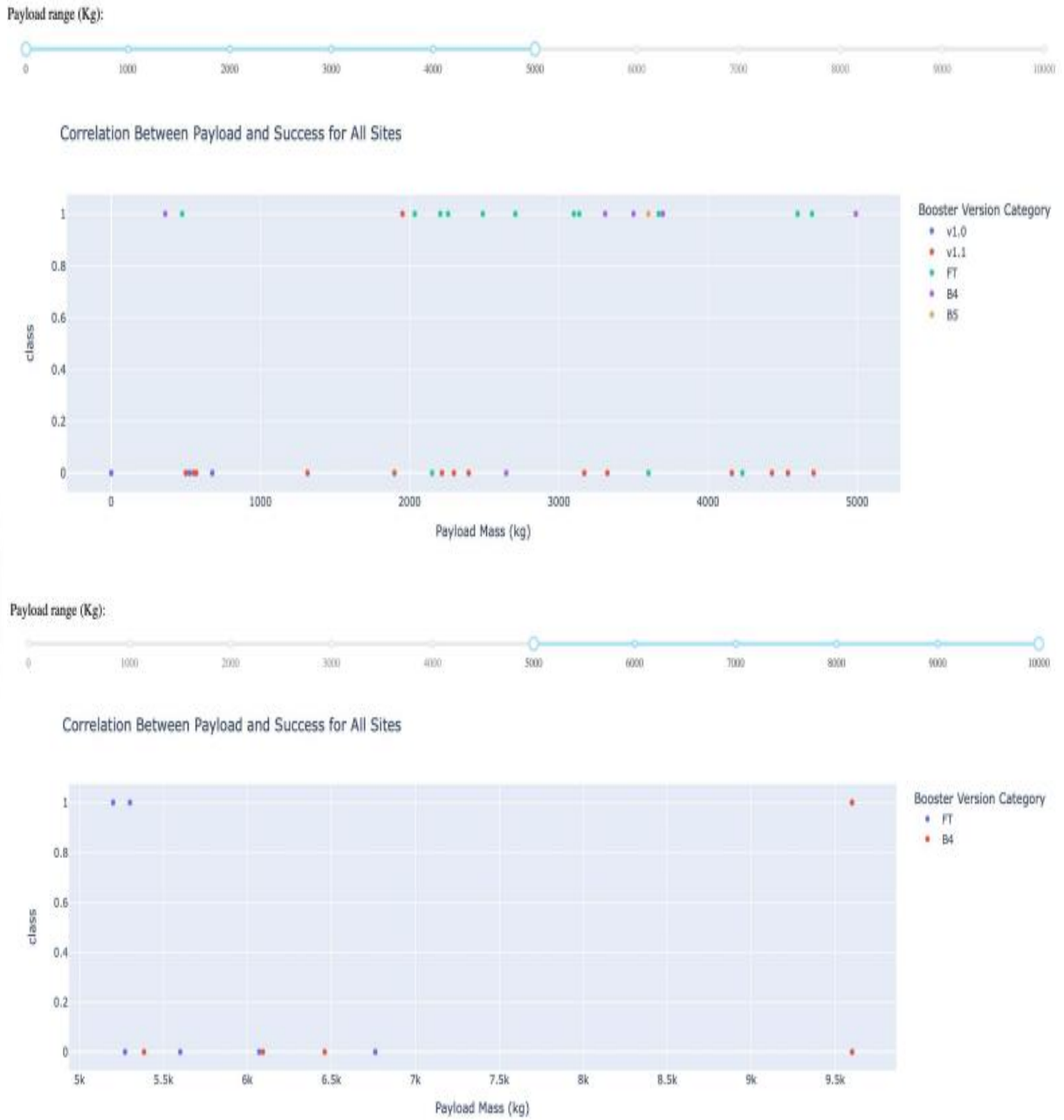
- **Explanation:**

• KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

- **Explanation:**

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.



Section 5

Predictive Analysis (Classification)

Out[52]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.933333	0.866667	0.866667	0.933333
F1_Score	0.965517	0.928571	0.928571	0.965517
Accuracy	0.944444	0.888889	0.944444	0.944444

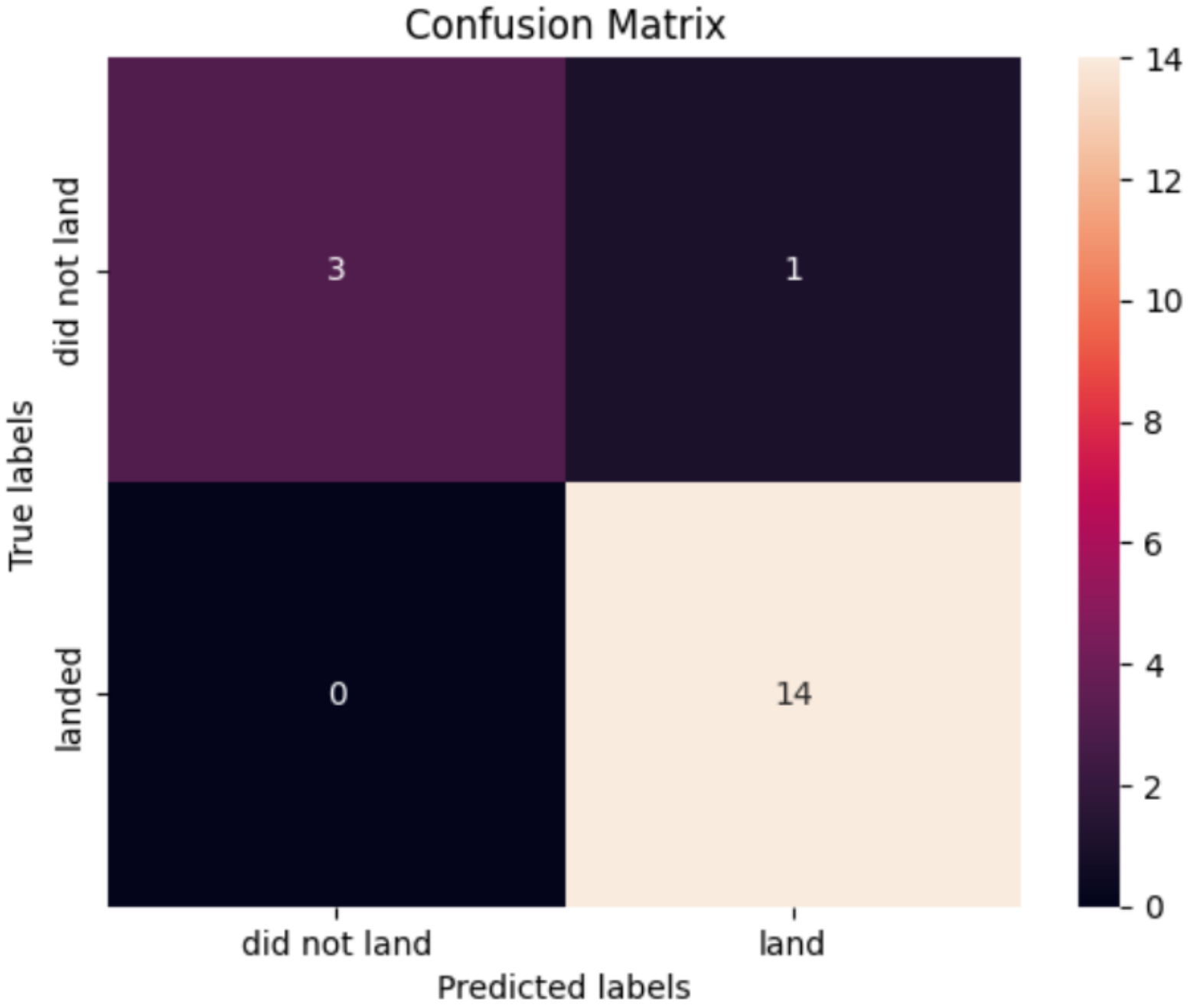
Classification Accuracy

- **Explanation:**

• Based on the scores of the Test Set, we can not confirm which method performs best.

• Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

• The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.



Confusion Matrix

- **Explanation:**
- • Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- The Decision Tree model stands out as the best algorithm for this dataset.
- Launches with lower payload masses tend to perform better compared to those with heavier payloads.
- Most launch sites are located near the Equator, and all of them are very close to the coast.
- The success rate of launches has been improving over the years.
- KSC LC-39A boasts the highest success rate among all the launch sites.
- Orbits such as ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!